1
Introdução

A *World Wide Web* (*WWW*) cresce desordenadamente, sem nenhum controle global ou planejamento, tornando-se estruturalmente complexa. Esse ambiente congrega milhões de participantes em tempo real, com interesses divergentes, criando e atualizando as estruturas dos *hyperlinks* (referência em um documento de hipertexto para outro documento) e, por conseguinte, inserindo novas informações e conhecimento.

Existem diversos trabalhos acerca das propriedades dos *hyperlinks* [28,29,30,31]. Por exemplo, há trabalhos que usam os *hyperlinks* para incrementar as classificações das máquinas de busca [11], melhorar os *web crawlers* [8,12], descobrir comunidades da *WWW* [10], fazer predições sobre a similaridade das páginas pesquisadas[13] e categorizar as páginas pelos seus *hyperlinks* [14, 15, 16, 17, 18].

Conjugando as características antes descritas, vem a lume um clássico problema na *WWW*: a busca de informações relevantes. A qualidade da busca depende, necessariamente, de um fator intrinsecamente subjetivo: a relevância. Tal fator advém de uma análise humana sobre a importância do conteúdo existente nas páginas que habitam o ambiente da *WWW*. Sendo assim, fez-se mister encontrar uma forma de coletar tal análise humana.

Tendo em vista tal necessidade, observou-se que o ambiente baseado em *hyperlink* possui na sua topologia informações substanciais sobre o seu conteúdo. A estrutura de *hyperlink* existente na *WWW*, constituído pelos *hyperlinks* das páginas, representa um considerável julgamento humano sobre a importância das páginas. A criação de um *hyperlink* na *WWW* representa uma indicação concreta de imputação de autoridade. Se a página p possui um *hyperlink* para uma página q, em algum momento o criador de p julgou importante, ou relevante, o conteúdo de q. Tal julgamento é essencial para formular o conceito de autoridade. Sendo assim, *hyperlinks* possibilitam encontrar potenciais autoridades avaliando, simplesmente, as páginas que

apontam para as mesmas.

Baseado nesse tipo de ambiente, *Jon Kleinberg* [1,2] desenvolveu um conjunto de algoritmos, chamado *HITS* (*Hyperlink Induced Topic Search*) [1,2,3,4,5,6], que utiliza a estrutura de *hyperlink* na *WWW* para extrair essas informações. Primeiramente, um sub-grafo dirigido da *WWW* é construído para uma consulta especifica. Em seguida, ele é expandido através dos *hyperlinks* existentes nas páginas do sub-grafo adicionando as páginas que possam exercer um papel de autoridade. Após a expansão do sub-grafo, um algoritmo iterativo processa a estrutura dos *hyperlinks* resultando numa extração e classificação de páginas relevantes para a busca.

Outros algoritmos para classificação de páginas, como *SALSA* e *PageRank*, são vastamente estudados em [5,7,19]. Adicionalmente, muitos trabalhos foram desenvolvidos baseados em variações do *HITS* [3,19].

O presente trabalho objetiva melhorar a qualidade dos resultados das buscas na *WWW* desenvolvendo uma nova tecnologia. Particularmente, essa nova tecnologia consiste em inserir novos conceitos ao modelo adotado por *Kleinberg*, resultando, pois, em uma extensão do modelo original.

Essa extensão adiciona novas considerações às definições, dadas por *Jon Kleinberg*, de *hubs* e autoridades: uma boa autoridade será uma página apontada por bons *hubs* e um bom *hub* será uma página que aponta para boas autoridades.

Nessa extensão as autoridades são apontadas por bons *hubs*, às vezes apontadas por bons portais e também apontam para boas novidades. Os bons *hubs* são páginas que apontam para boas autoridades e novidades, e são apontados por bons portais. As boas novidades são páginas que são apontadas pelas boas autoridades, pelos bons *hubs* e pelos bons portais e bons portais são páginas que apontam para as boas autoridades, para bons *hubs* e para boas novidades. Assim, formulamos um algoritmo estendido, *XHITS (Extended Hyperlink Induced Topic Search),* que visa melhorar a classificação das autoridades do ambiente.

A ênfase da presente dissertação é definir essa nova extensão,

mostrando a sua influência no cálculo das novas autoridades. Em seguida, através de um ambiente de experimentação desenvolvido especialmente para esse trabalho, passamos a coletar classificações utilizando o *HITS* e o *XHITS* e, por conseguinte, subtrair dados comparativos sobre as suas classificações. Por fim, em conclusão, demonstramos a proficiência do novo algoritmo *XHITS,* no que se refere à classificação das páginas apresentadas, quando comparado com o algoritmo *HITS*. Ele apresenta uma melhora em torno de setenta por cento. Para medir esse ganho adotamos duas referencias, a saber: o Google e o Yahoo. Estas duas máquinas de busca funcionam como especialistas artificiais que fornecem a baixíssimo custo as respostas que para efeitos de teste consideramos as corretas.

Na seção 2 desse trabalho detalhamos o modelo de *Hubs* e *Authorities* proposto por *Jon Kleinberg*. Em seguida, na seção 3, descrevemos a extensão do modelo de *Hubs e Authorities*. Na seção 4, demonstramos, especificamente, o ambiente de experimentação desenvolvido e, ainda, descrevemos como são efetuadas as comparações dos resultados. Na seção 5, apresentamos os resultados obtidos e as análises correspondentes e, por fim, a conclusão dessa dissertação.